USDA

United States
Department of
Agriculture

A Compilation of PEDITOR Estimation Formulas

Charles Day

National
Agricultural
Statistics
Service

Research and
Development
Division

RDD Research Report
Number RDD-02-03

January 2002

A COMPILATION OF PEDITOR ESTIMATION FORMULAS, compiled by Charles Day, Research and Development Division, National Agricultural Statistics Service, U. S. Department of Agriculture, Washington, DC, January 2002, NASS Research Report No. RDD-02-03.

ABSTRACT

This report provides a handy "one-stop" reference for all of the estimation formulas used in NASS's PEDITOR remote sensing image processing and estimation software. It is intended as meaningful documentation for the Agency's remote sensing analysts in the State Statistical Offices and in headquarters.

KEYWORDS

Remote sensing; estimation; county estimates; PEDITOR.

ACKNOWLEDGMENTS

This work consists entirely of a compilation of the work of others. In particular, the author wishes to thank Bill Wigton, Don Von Steen, Richard Sigman, Gail Walker, Michael Bellow, Michael Craig, and George Hanuschak for their original efforts in developing and extending the methodology described here, and Robert Hale for his earlier effort to gather much of this material in an internal document.

TABLE OF CONTENTS

INTRODUCTION

PEDITOR is NASS's internal image processing and estimation package for use with satellite remote sensing data. This short paper consolidates the estimation formulas used in the various programs in PEDITOR.

Satellite images are composed of pixels (or picture elements) much like the image on a computer monitor. The satellites being used in our acreage estimation work are all equipped with sensors which collect electromagnetic (EM) energy in several bands of the EM spectrum. Each pixel in the image is an n-tuple, consisting of one observation from each of the n sensors.

In order to do estimation, a "ground truth" sample is needed; that is, a sample of areas where acreages and cover types are known. Fortunately, NASS already has its area frame sample which meets these criteria. The area frame is constructed by dividing each state into "primary sampling units" (PSUs). These units are evaluated using satellite and photographic imagery, and each unit is assigned to a sampling stratum based on the proportion of land in use for agricultural activity. A stratified systematic sample is drawn, and, within each selected primary sampling unit, "segments" (smaller divisions of uniform size based on the stratum) are drawn off. For each selected PSU, a segment is then selected, and a NASS enumerator is sent to draw off field boundaries, determine what crop or other land covers are in the fields, and the acreage of each field. The remote sensing program uses this information to help train a maximum likelihood classifier, which can then be used to classify all of the pixels in an entire satellite scene.

For purposes of estimation, NASS divides the region of interest, usually a state or part of a state, into "Analysis Districts." An Analysis District is defined as, "a unique area of land to be analyzed by a separate analysis. Analysis Districts are characterized by the same date(s) of [satellite] imagery or as an area having no satellite coverage, but included in the original region of interest," (Craig, unpublished training material). Analysis Districts are built up by aggregating "subcounties." A subcounty is defined as, "a specific part of a county or parish that is wholly contained in a given, selected [satellite] scene." Note that, using this definition, a subcounty may be (and very often is) a whole county. In all cases, state level estimates are made by aggregating analysis district estimates.

There are five estimation methods currently available in PEDITOR, two of which were added for the 2000 crop season. The decision as to which method will be used is made at the level of sampling stratum within Analysis District. The best, and most frequent, situation is that an Analysis District has cloud-free satellite imagery available from dates during the growing season, and the stratum has sufficient ground truth for a valid regression to be performed. In this case, regression estimation, with the pixel counts classified to a particular crop cover serving as the auxiliary variable, and the observed number of acres of a crop cover from the area frame survey as the variable of interest is

recommended. If the regression estimation methods are used, then the Battese-Fuller county estimation method is used to make estimates at the county level.

Two alternatives are available if the data do not support the use of a separate regression for each stratum. The first is combined regression, in which two or more strata are combined for purposes of making a regression estimate. This method is older, and is no longer recommended, for reasons discussed below. The recommended method when insufficient data are available for separate regression is the Simple Adjusted Pixel Count Estimator (SAPCE). This method is also discussed below. While the decision is made to use this estimation method at the Analysis District/stratum level, this estimation method itself is performed at the subcounty level. County and Analysis District level estimates are made by aggregating the subcounty estimates.

When there are no satellite data available for an Analysis District, or, in the rare event that the available imagery does not yield a usable classification, the weighted and unweighted proration methods are available. These methods prorate the June Agricultural Survey (JAS) area frame estimate for the crop of interest to each subcounty based on the number of area frame sampling units in the subcounty. Again, county and Analysis District estimates are made by aggregation. The weighted method is recommended; it uses the previous 3 years' county estimates for the crop of interest to help allocate the JAS estimates properly by county, rather than assuming a uniform allocation. The unweighted method is only used when no prior years' county estimate information is available.

The chart in Appendix 1 summarizes the procedure for choosing which estimator to use.

Unless otherwise noted, all quantities in the estimation formulas below refer to a particular crop cover within an analysis district. Subscripts indicating crop cover and analysis district are omitted to simplify the notation.

SEPARATE REGRESSION ESTIMATOR FOR ACRES OF THE CROP COVER OF INTEREST IN STRATUM h

The separate regression estimator for number of acres of the crop cover of interest in a single stratum h is:

$$\hat{y}_h = N_h\left[\overline{y}_h + b_h(\overline{X}_h - \overline{x}_h)\right]$$

where:

N_h = The number of frame units (segments in the frame) in stratum h

N = The number of frame units in all strata

$\bar{y}_h$ = The (sample) mean (per segment) of reported acres of the crop cover of interest in stratum h

b_h = The slope from the regression of number of acres (of the crop cover of interest in a segment) on number of pixels (classified to that crop cover in the segment) in stratum h

$\bar{X}_h$ = The population mean number of pixels in a segment classified to the crop cover of interest in stratum h

$\bar{x}_h$ = The sample mean number of pixels in a segment classified to the crop cover of interest in stratum h

Note that this estimator, developed by Von Steen and Wigton (1976), uses the remote sensing data about number of pixels classified to a particular crop cover as an auxiliary variable. Note further that $(\bar{X} - \bar{x})$ is the difference between the mean number of pixels classified to the crop cover of interest in a segment in the population, and the mean number of pixels classified to the crop cover of interest in a sampled (training) segment. Since **b** converts pixels to acres, $b(\bar{X} - \bar{x})$ is the average difference in acres classified to the crop cover of interest between a population segment and a sampled segment. This is used to adjust the sample mean number of acres in a sampled segment before multiplying by the number of segments in the analysis district to get an estimate of the total number of acres of the crop cover of interest in that analysis district.

It is a rule of thumb, based on a simulation done by Chhikara and McKeon (1986), that a stratum should have ten or more observations in order for the variance to be estimated with an acceptably small error. If there are fewer than ten observations in a stratum, then the analyst should consider using the Simple Adjusted Pixel Count Estimator (SAPCE) or Combined Regression Estimator.

VARIANCE OF $\hat{y}_h$ (SEPARATE REGRESSION ESTIMATOR FOR ACRES OF CROP COVER OF INTEREST IN STRATUM h)

The formula for the estimator of the variance of the single stratum (not combined) regression estimate is:

$$\mathrm{var}\left(\hat{y}_h\right) = (N_h^2 / n_h)(1 - f)\left[\sum_{i \in H} (y_i - \bar{y}_h)^2 / (n_h - 2)\right](1 - R_h^2)\left[1 + (1/(n_h - 3)\right]$$

where: $f = n_h / N_h$

$$R_h^2 = (S_{xy_h}^2)^2 / (S_{y_h}^2 \cdot S_{x_h}^2)$$

$$S_{xy_h}^2 = \sum_{i \in H} (x_i - \bar{x}_h)(y_i - \bar{y}_h) / (n_h - 1) =$$

$$(\sum_{i \in H} x_i y_i - n_h \bar{x}_h \bar{y}_h) / (n_h - 1)$$

$$S_{y_h}^2 = \sum_{i \in H} (y_i - \bar{y}_h)^2 / (n_h - 1) = (\sum_{i \in H} y_i^2 - n_h \bar{y}_h^2) / (n_h - 1)$$

$$S_{x_h}^2 = \sum_{i \in H} (x_i - \bar{x}_h)^2 / (n_h - 1) = (\sum_{i \in H} x_i^2 - n_h \bar{x}_h^2) / (n_h - 1)$$

H is the set of segments in stratum h with the crop cover of interest.

Note that this is equivalent to the variance estimator shown in Cochran (*Sampling Techniques*, 2nd edition, p. 202), with an approximate adjustment factor $[1 + (1 / n_h - 3)]$ suggested by Cochran ("Sampling Theory When the Sampling-Units are of Unequal Sizes," *JASA*, Vol. 37, 1942, pp. 199-212) to account for the fact that the segments are of unequal size. Note also that as R^2 approaches one, the variance approaches zero, implying that strata with strong linear relationships between number of pixels classified to a cover and number of acres of that cover will get the greatest improvement in precision over the direct expansion estimator. In fact, in practice, the average reduction in variance for major crops in the states selected for this program has been in the 80 to 90 percent range. Note further that, in this paper, "Var" will be used to designate a variance, while "var" will be used to designate a variance estimator.

COMBINED REGRESSION ESTIMATOR

Strata should only be combined when they have similar land use stratification and the same target segment size. For example, a stratum with greater than 75 percent agricultural land might be combined with a stratum with between 50 and 75 percent agricultural land, but neither would ever be combined with a urban or woodland stratum. The combined estimator is appropriate when it is reasonable to believe that the true regression coefficients are equal in all of the strata being combined. In particular, it should be reasonable to believe that the classification is working about equally well in all of the strata being combined.

The burden of these assumptions is not easy to meet. Further, the combined regression coefficient is known to be biased, with a bias on the order 1/n. In past years, when the only alternative was the unweighted proration estimator, which has its own strong assumptions to be met and its own practical problems, combined regression was considered the first alternative when there was no valid separate regression in a stratum. Now, with the SAPCE estimator available, combined regression should be used infrequently, and only then if there is strong evidence that its assumptions are met. The combined estimator for strata with two or more observations is:

$$\hat{y}_c = \sum_{h \in C} N_h \left[\bar{y}_h + b_c \left(\bar{X}_h - \bar{x}_h \right) \right]$$

where the differences from the single stratum estimator are:

$$b_c = \left(\sum_{h \in C} a_h \cdot S^2_{xy_h} \right) / \left(\sum_{h \in C} a_h \cdot S^2_{x_h} \right)$$

$$a_h = \left(N_h^2 / n_h \right) \cdot \left(1 - \left(n_h / N_h \right) \right)$$

Note that this estimate of b_c is not the pooled estimate. The pooled estimate would require additional assumptions to hold in order to be valid.

C is the set of strata over which the combined estimate is being made.

H is the set of segments in stratum h with the crop cover of interest.

The following changes must be made in the calculations for strata that are to be combined but have fewer than 2 segments:

$\bar{y}_h$ Must be replaced with the weighted mean (weighted by number of frame units) of the $\bar{y}_h$ in the strata that do have 2 or more segments.

$\bar{x}_h$ Must be replaced with the weighted mean (weighted by number of frame units) of the $\bar{x}_h$ in the strata that do have 2 or more segments.

b_c Strata with fewer than 2 segments should be excluded from the model for developing the slope.

VARIANCE OF $\hat{y}_c$ (COMBINED ESTIMATOR (TWO OR MORE STRATA) FOR ACRES OF CROP COVER OF INTEREST) AND ESTIMATE OF R^2 FOR THE COMBINED REGRESSION

When each of the strata has two or more segments, the estimated variance of the combined estimate is given by:

$$\text{var}(\hat{y}_c) = \sum_{h \in C} [a_h \cdot s_h^2 \cdot (1 + (2/(n - k - 2)))]$$

where:

$$a_h = (N_h^2 / n_h) \cdot (1 - (n_h / N_h))$$

$$s_h^2 = \sum_{i \in H} \left[(y_i - \bar{y}_h) - b_c (x_i - \bar{x}_h) \right]^2 / (n_h - 1)$$

$$b_c = \left(\sum_{h \in C} a_h \cdot S_{xy_h}^2\right) / \left(\sum_{h \in C} a_h \cdot S_{x_h}^2\right)$$

$$n = \sum_{h \in C} n_h$$

H is the set of segments in stratum h with the crop cover of interest.

k is the number of strata being combined.

Note the change in the adjustment factor from $[1 + (1/(n_h - 3))]$ to $[1 + (2/(n - k - 2))]$. Besides the obvious adjustment for number of strata and the use of the combined n, there is an additional adjustment for the degree of freedom lost in the estimation of the combined regression coefficient which accounts for the 2 in the numerator of the fractional part of the adjustment factor (Chhikara and McKeon, 1986, p. 3). Note that if two strata with two observations each were combined, $n - k - 2$ would equal zero. This is not a problem in practice, since a regression estimate would not be attempted for such a combination of strata. (The total number of segments with the crop cover of interest in the combined strata is too small to make the regression estimate practical.)

If one or more of the strata has fewer than two segments with the crop cover of interest, the variance estimate is given by:

$$\mathrm{var}(\hat{y}_c) = [1 + \sum_{h' \in H'} ((2N_{h'} / (\sum_{h'' \in H''} N_{h''})) + (N_{h'} / (\sum_{h'' \in H''} N_{h''}))^2)] \cdot \mathrm{var}_{H''}$$

where:

$$\mathrm{var}_{H''} = \sum_{h'' \in H''} [a_{h''} \cdot s^2_{h''} \cdot (1 + (2/(n - k - 2)))]$$

H'' is the set of all strata containing two or more segments
H' is the set of all strata containing fewer than two segments

Note that this amounts to computing the combined variance as we did before for the strata with two or more segments. The variance for strata with fewer than 2 segments is then computed by applying a weighting factor based on the proportion of frame units in strata with fewer than two segments per stratum to frame units in strata with two or more segments. These two pieces of the variance are then summed to yield the total variance of the combined estimate.

An estimate of R^2 for the combined regression, denoted R^2_c may be made using the estimated variance of the direct expansion estimate, denoted Var(DE), and the estimated variance of the regression estimate, denoted Var(REG), by means of the following equation:

$$R^2_c = [\mathrm{Var(DE)}_c - \mathrm{Var(REG)}_c] / \mathrm{Var(DE)}_c$$

where:

$\mathrm{Var(REG)}_c = \mathrm{Var}(\hat{y}_c)$, computed as appropriate, depending on whether or not any stratum has fewer than two segments.

$$\mathrm{Var(DE)}_c = [\sum_{h'' \in H''} \mathrm{Var}_{h''}] \cdot [1 + ((\sum_{h' \in H'} N_{h'}) / (\sum_{h'' \in H''} N_{h''}))]^2$$

where

$\mathrm{Var}_{h''}$ is the variance in stratum h'' of the NASS area frame direct expansion (DE) estimator.

Note that this is effectively a weighting up of the direct expansion estimator for strata which contain two or more segments with the crop cover of interest to account for the frame units in the strata which contain fewer than two such segments.

This formula for R^2 takes advantage of the relationship between the variance estimator for the regression estimate and the variance estimator for the direct expansion estimate. A brief examination of the variance estimator for the one stratum regression estimate shows that, the regression variance estimator is approximately (ignoring adjustment factors) $(1-R^2)$ times the direct expansion variance estimator. A little manipulation of that relationship yields the above formula for R^2.

SIMPLE ADJUSTED PIXEL COUNT ESTIMATOR

Occasionally, the situation occurs that, because of the number of segments with a cover of interest in a particular stratum in an analysis district, neither the separate nor combined regression estimators is appropriate, yet the classification for that analysis district is of good quality. In these cases, NASS uses an estimator based on simply counting the number of pixels classified to the cover of interest in that analysis district. This is the Simple Adjusted Pixel Count Estimator (SAPCE).

Some additional assumptions and notation are required:

X_{ihk} = number of pixels classified to desired cover type in stratum h, subcounty k of analysis district i

$X_{i..}$ = number of pixels classified to desired cover type in analysis district i (across all strata and subcounties)

λ = conversion factor (areal units per pixel)

m_{ilt} = total number of sample pixels in analysis district i labeled cover type "l" in the ground truth and classified to cover type "t". Note that this number is across all segments in the analysis district, and is not subcounty or stratum specific.

Then

$m_{ip.}$= the marginal total of all sample pixels labeled cover "p" (the desired cover type)

and

$m_{i.p}$ is the marginal total of all sample pixels categorized to cover "p".

Then the Simple Adjusted Pixel Count Estimator (SAPCE) for desired crop/cover type "p", subcounty k, and stratum h of analysis district i is:

$$S_{ihk} = \lambda (m_{ip.}/m_{i.p}) X_{ihk}$$

The new SAPCE estimator for the entire analysis district i is:

$$S_{i..} = \sum_{h} \sum_{k \in AD_i} S_{ihk}$$

The new SAPCE estimator for whole subcounty c is:

$$S_{...(c)} = \sum_{i} \sum_{h} \sum_{k \in \text{subcounty "c"}} S_{ihk}$$

In order to calculate the variance of S_{ihk}, a jackknife approach is used. In this approach, one segment is dropped out and the ratio $m_{ip.}/m_{i.p}$ is recalculated based on the new data set. If we define:

$n_{i.}$ = number of sampled segments used to create signatures for classification. (Because of overlap at the edges of the satellite scenes, a segment may be contained in more than one scene. When analysis districts are defined, each segment is defined as being in only one analysis district; however, all of the segments in a scene, regardless of which analysis district they belong to, are used for creating signatures. So n_i contains sampled segments which lie in the overlap between the scenes used in this analysis district and scenes used in adjacent analysis districts which are defined as being in the adjacent analysis districts.)

$m_{ip.(s)}$ = recalculated $m_{ip.}$ after deleting segment s from analysis district i

$m_{i.p(s)}$ = recalculated $m_{i.p}$ after deleting segment s from analysis district i

$K_{is} = m_{ip.(s)}/m_{i.p(s)}$, where s is the segment dropped out.

Then the variance of $m_{ip.}/m_{i.p}$ is given by:

$$Var\left(\frac{m_{ip.}}{m_{i.p}}\right) = \frac{(n_{i.}-1)}{n_{i.}} \sum_{s}^{n_i} \left(K_{is} - \frac{m_{ip.}}{m_{i.p}}\right)^2$$

An estimate of the variance of the desired crop/cover type, subcounty k, and stratum h of analysis district i is:

$$\text{var}(S_{ihk}) = \text{var}\left(\frac{m_{ip.}}{m_{i.p}}\right) \cdot [\lambda X_{ihk}]^2$$

and the estimated variance for $S_{i...}$, the pixel estimate for the analysis district i, is given by:

$$\text{var}(S_{i..}) = \text{var}\left(\frac{m_{ip.}}{m_{i.p}}\right) \cdot [\sum_{h} \sum_{k \in AD_i} \lambda X_{ihk}]^2$$

and the variance of the entire county estimate $S_{...(c)}$ is:

$$\text{var}(S...(c)) = \sum_{i} \text{var}\left(\frac{m_{ip.}}{m_{i.p}}\right) [\sum_{h} \sum_{k \in \text{County "c"}} \lambda X_{ihk}]^2$$

These variance calculations maintain a constant coefficient of variation (CV) for the estimate when any parts of an analysis district are summed (by county or by county and strata to get analysis district). Within the analysis district, the CV of any acreage estimate is always kept equal to the CV of the jackknifed variable:

$$\text{cv}(S_{i..}) = \text{cv}\left(\frac{m_{ip.}}{m_{i.p}}\right) = \frac{\sqrt{\text{var}\left(\frac{m_{ip.}}{m_{ip.}}\right)}}{\frac{m_{ip.}}{m_{i.p}}}$$

ESTIMATION WITH CLOUD COVER OR IN THE ABSENCE OF AN ACCEPTABLE CLASSIFICATION

One of the problems with estimation of crop areas with satellite data is the use of imagery which contains clouds. The satellite depends on reflected energy in the visible and infrared parts of the electromagnetic spectrum to record its observations. When a particular area is covered by clouds, the reflected energy from the clouds, rather than the ground, is recorded. As a result, there are no classified pixel counts available for crops in the cloud covered area, and the cloud covered areas cannot be estimated in the usual way. One might suggest that the cloud covered areas could be treated as occurring at random. This was, in fact, the assumption of the interdepartmental Large Area Crop Inventory

Experiment (LACIE) project. Research showed that this assumption was of questionable validity. Intensive crop growth was, of course, associated with areas of greater rainfall, and thus with areas more likely to be covered by clouds. The follow-on AGRISTARS project recognized the need for a method to make estimates for these cloud-covered areas.

There are rare occasions when serious problems with the crop cover type classification of the satellite pixels may occur, despite the fact that there is cloud-free imagery. This may occur when the available dates of cloud-free imagery fall too close to the beginning or end of the growing season for different cover types to be properly differentiated, or if there is a dearth of ground truth for one or more cover types. An estimation method was required that utilized the June area sample ground data for this domain.

The weighted and unweighted proration methods described below are used in these situations. The weighted method was developed by Bellow (1994) and Craig (forthcoming); the unweighted method was developed by Hanuschak (1976). The unweighted method has been in use for many years, and was initially designed primarily for state-level estimation. It's assumptions are not as likely to hold if applied to domains (such as counties). The unweighted method assumes that the distribution of crops across the subcounties in an analysis district is the same. In practice, violation of this assumption has sometimes resulted in positive estimates for crops in some counties where the crop is known not to be grown. For that reason, the weighted proration estimator was developed. The weighted estimator uses a ratio of the previous 3 years' average estimate for each county to the total estimate for the state in order to apportion crops only to counties in which they are being grown. These newer methods, due to Craig, have resulted in improved county-level estimates.

Unweighted Proration

Consider the analysis district to be the union of two domains, the cloud-free domain, and the cloud-covered domain. (Treat these domains as post-strata.)

Let: j = 1 represent the cloud-free domain
j = 2 represent the cloud-covered domain

then y'_{jhi} is defined as the number of acres of the crop cover being estimated in domain j, stratum h, and segment i.

The total estimator for the cloud covered domain is:

$$\hat{Y}_2 = \sum_{h=1}^{L} N_h \left(\sum_{i=1}^{n_h} y'_{2hi} \right) / n_h$$

This is the "direct expansion" estimator applied to the segments in the cloud covered area. The associated variance estimator is:

$$\text{var}(\hat{Y}_2) = \sum_{h=1}^{L} (N_h^2 / (n_h(n_n - 1)))((N_h - n_h) / N_h) \cdot$$

$$[\sum_{i=1}^{n_h} y_{2hi}^{'2} - ((\sum_{i=1}^{n_h} y_{2hi}^{'})^2 / n_h)]$$

The total for the cloud-free domain is estimated in the usual way, using only the segments in the cloud-free domain. The total estimator for the cloud-free domain is:

$$\hat{Y}_1 = \sum_{h=1}^{L} N_{h1} y_{h.}^{'}$$

where

$$y_{1h.}^{'} = \overline{y}_{1h} + b_h(\overline{X}_{1h} - \overline{x}_{1h})$$

$\overline{y}_{1h}$ = average number of acres per sample segment of the crop cover being estimated in stratum h in the cloud-free domain

$\overline{X}_{1h}$ = average number of pixels of the crop cover being estimated per segment in stratum h in the entire cloud-free domain

$\overline{x}_{1h}$ = average number of pixels of the crop cover being estimated per segment in the ground truth sample in the cloud-free domain

The associated variance estimator is:

$$\text{var}(\hat{Y}_1) = \sum_{h=1}^{L} (N_h^2 / n_h)((N_h - n_h) / N_h)[\sum_{i=1}^{n_h} y_{1hi}^{'2} - ((\sum_{i=1}^{n_h} y_{1hi}^{'})^2 / n_h)] \cdot$$

$$[(1 - R_h^2) / (n_h - 2)]$$

To obtain the estimate for the whole analysis district, simply add the estimates for the two domains:

$$\hat{Y} = \hat{Y}_1 + \hat{Y}_2$$

The variance is obtained by the usual formula for the sum of two nonindependent random variables:

$$\mathrm{Var}(\hat{Y}) = \mathrm{Var}(\hat{Y}_1) + \mathrm{Var}(\hat{Y}_2) + 2\mathrm{Cov}(\hat{Y}_1, \hat{Y}_2)$$

with the following covariance term:

$$\mathrm{cov}(\hat{Y}_1, \hat{Y}_2) = \sum_{h=1}^{L} W_h^2 \, \mathrm{cov}(\hat{Y}_{1h}, \hat{Y}_{2h})$$

where:

$$\mathrm{cov}(\hat{Y}_{1h}, \hat{Y}_{2h}) = -N_h^2[(\sum_{i=1}^{n_h} y_{1hi}^{\cdot})(\sum_{i=1}^{n_h} y_{2hi}^{\cdot})] / (n_h(n_h - 1)),$$

and

$$W_h = \frac{N_h}{N}$$

Weighted Proration

Assume:

i = Analysis District
j = stratum
k = unique county or subcounty
c = original whole county, associated with a unique subcounty k above
s = segment
y_{js} = total acres of crop cover type of interest for segment s in stratum j
N_{jk} = number of frame units in subcounty k and stratum j
$N_{j(c)}$ = number of frame units in original county "c", stratum j (i.e., sum of all k's $\in$ county c)
$N_{j.}$ = number of frame units in the entire state in stratum j
n_j = number of sample segments in stratum j (across all analysis districts)
w_c = weight (average of the previous 3 year's State Statistical Office estimates for the crop of interest, county c)
w. = sum of the w_c across all counties in state

Then, if we define:

JAS_j = current year June Agricultural Survey direct expansion estimate for stratum j,

$$JAS_j = N_{j..}(\sum_{s=1}^{n_j} y_{js}) / n_j$$

and

$Var(JAS_j)$ = variance of JAS_j,

$$Var(JAS_j) = \frac{N_{j..}^2}{n_j(n_j - 1)} \frac{N_{j..} - n_j}{N_{j..}} \left(\sum_{s=1}^{n_j} y_{js}^2 - \frac{(\sum_{s}^{n_j} y_{js})^2}{n_j}\right)$$

and

$$R_c = w_c / w_.$$

and finally, define the subcounty part estimate (where each k is associated with a county c) to be:

$$M_{jk} = (N_{jk} / N_{j(c)}) \cdot R_c \cdot (JAS_j)$$

(Note: if $N_{j(c)}$=0; then set N_{jk} = 1 for all k's part of county c, and set $N_{j(c)}$ = number of k's)

The weighted proration estimator for Analysis District i is:

$$A_i = \sum_j \sum_{k \in AD_i} M_{jk}$$

and the subcounty variance estimate (a proration of the overall variance) is given by:

$$var(M_{jk}) = (N_{jk} / N_{j(c)}) \cdot (R_c)^2 \cdot var(JAS_j)$$

then the overall estimated variance for analysis district i is:

$$var(A_i) = \sum_j \sum_{k \in AD_i} var(M_{jk})$$

The county 'c' weighted proration estimator is:

$$T_c = \sum_j \sum_{k \in \text{County "c"}} (M_{jk})$$

and the overall county variance estimate is given by:

$$\text{var}(T_c) = \sum_j \sum_{k \in \text{County "c"}} \text{var}(M_{jk})$$

COUNTY LEVEL REGRESSION ESTIMATION PROCEDURES IN PEDITOR

County level regression estimates are made in PEDITOR using the Battese-Fuller method described by Walker and Sigman (1982). To determine the county level estimate for a county "c" using the state level regression estimate, the number of frame units in a county are multiplied times the adjusted county mean. **NOTE:** The subscript "c" in this section refers to *county*, not to a combined estimator as in the previous section.

$$\hat{Y}_{hc} = N_{hc} \cdot \overline{Y}_{\delta}$$

where:

$$\overline{Y}_{\delta} = (1-\delta)\overline{Y}_0 + \delta \overline{Y}_1$$

$$\overline{Y}_1 = \overline{Y}_c + b_{hAD}(\overline{X}_c - \overline{x}_c)$$

$$\overline{Y}_0 = b_{1hAD}\overline{X} + b_{0hAD}$$

b_{0hAD} = analysis district intercept by stratum
b_{1hAD} = analysis district slope by stratum

where $\overline{Y}_c$, $\overline{X}_c$, and $\overline{x}_c$ are the subcounty mean reported acres, the population pixel mean in the subcounty, and the sample pixel mean in the subcounty respectively for the crop of interest in stratum h. That is, they are the subcounty level analogs to the similar variables in the Analysis District level estimator.

The following five rules, in order of precedence, determine the proper δ value to use:

1) Use $\delta = 1$ iff $\sigma^2_{within} = 0$.

2) If $\sigma^2_{between} = 0$, use $\delta = 0$.

3) If no county in the analysis district has more than 2 segments use $\delta = 0$.

4) If $\sigma^2_{within} = 1.0$, use $\delta = 0$

5) otherwise use $\delta = \Gamma$, which is the value which minimizes MSE

where:

$\Gamma = \sigma^2_{between} / (\sigma^2_{between} + \sigma^2_{within}/n)$ where:

$\sigma^2_{between}$ = The variance between county means within an analysis district by stratum

σ^2_{within} = The variance of reported data within a county by stratum

The variance for the county level estimate is

$Var_{hc} = \sigma^2_{between} + \sigma^2_{within}$

Bellow (1994) gives the following estimators for the Battese-Fuller variance components:

$$\hat{\sigma}^2_{within} = [1/(n_h - C - 1)] \sum_{c=1}^{C} \sum_{i=1}^{n_{hcty}} [y_{hci} - \bar{y}_{hc.} - \hat{\alpha}_h (x_{hci} - \bar{x}_{hc.})]^2$$

$$\hat{\sigma}^2_{between} = \max[0, (s^2_{uh} - (n_h - 2)\hat{\alpha}^2_{eh}) / (n_h - T_h)]$$

where:

$$\hat{\alpha}_h = [\sum_{c=1}^{C} \sum_{i=1}^{n_{hcty}} (x_{hci} - \bar{x}_{hc.})(y_{hci} - \bar{y}_{hc.})] / \sum_{c=1}^{C} \sum_{i=1}^{n_{hc}} (x_{hci} - \bar{x}_{hc.})^2$$

$$s^2_{uh} = \sum_{c=1}^{C} \sum_{i=1}^{n_{hc}} (y_{hci} - \hat{\beta}_{0h} - \hat{\beta}_{1h} x_{hci})^2$$

$$T_h = n_h \sum_{c=1}^{C} n_{hc.}^2 \bar{x}_{hc.}^2 + (\sum_{c=1}^{C} n_{hc}^2)(\sum_{c=1}^{C} \sum_{i=1}^{n_{hc}} x_{hci}^2) - 2n_h \bar{x}_{h..} \sum_{c=1}^{C} n_{hc}^2 \bar{x}_{hc} /$$

$$(n_h \sum_{c=1}^{C} \sum_{i=1}^{n_{hc}} x_{hci}^2) - n_h^2 \bar{x}_{h..}^2$$

PUTTING IT ALL TOGETHER

Much of the estimation process in PEDITOR has been automated using the RESTP module. This module guides the analyst through the process of making estimates. As a part of its functioning, it determines for each crop, Analysis District, and area frame stratum which estimator to use in the following order of preference: Regression, Pixel Count (Simple Adjusted Pixel Count by default), Weighted Proration, and Unweighted Proration. The proration, weighted proration, and pixel count estimates are calculated at the subcounty/stratum level and aggregated to the analysis district/stratum level. The regression estimates are made at the analysis/district stratum level. For county estimation, separate regression estimates are made using the county level regression estimation procedure outlined above for each subcounty where regression is to be used. . County estimates are made by aggregating the appropriate subcounty/stratum estimates for each county for each crop. State estimates are made by aggregating the Analysis District/Stratum level estimates for each crop.

REFERENCES

Bellow, M. E., *Application of Satellite Data to Crop Area Estimation at the County Level,* U. S. Department of Agriculture, NASS Research Report No. STB-94-02, 1994.

Battese, G. E., Harter, R.M., and Fuller, W. A., "An Error-Components Model for Prediction of County Crop Areas using Survey and Satellite Data," *Journal of the American Statistical Association,* Volume 83, No. 401, 1988, pp. 28-36.

Chhikara, R. S., and McKeon, J. J., *Estimation of County Crop Acreages Using Landsat Data as Auxiliary Information,* Houston, Texas: University of Houston, unpublished.

Cochran, W. G., *Sampling Techniques,* John Wiley and Sons, 2nd ed., 1963.

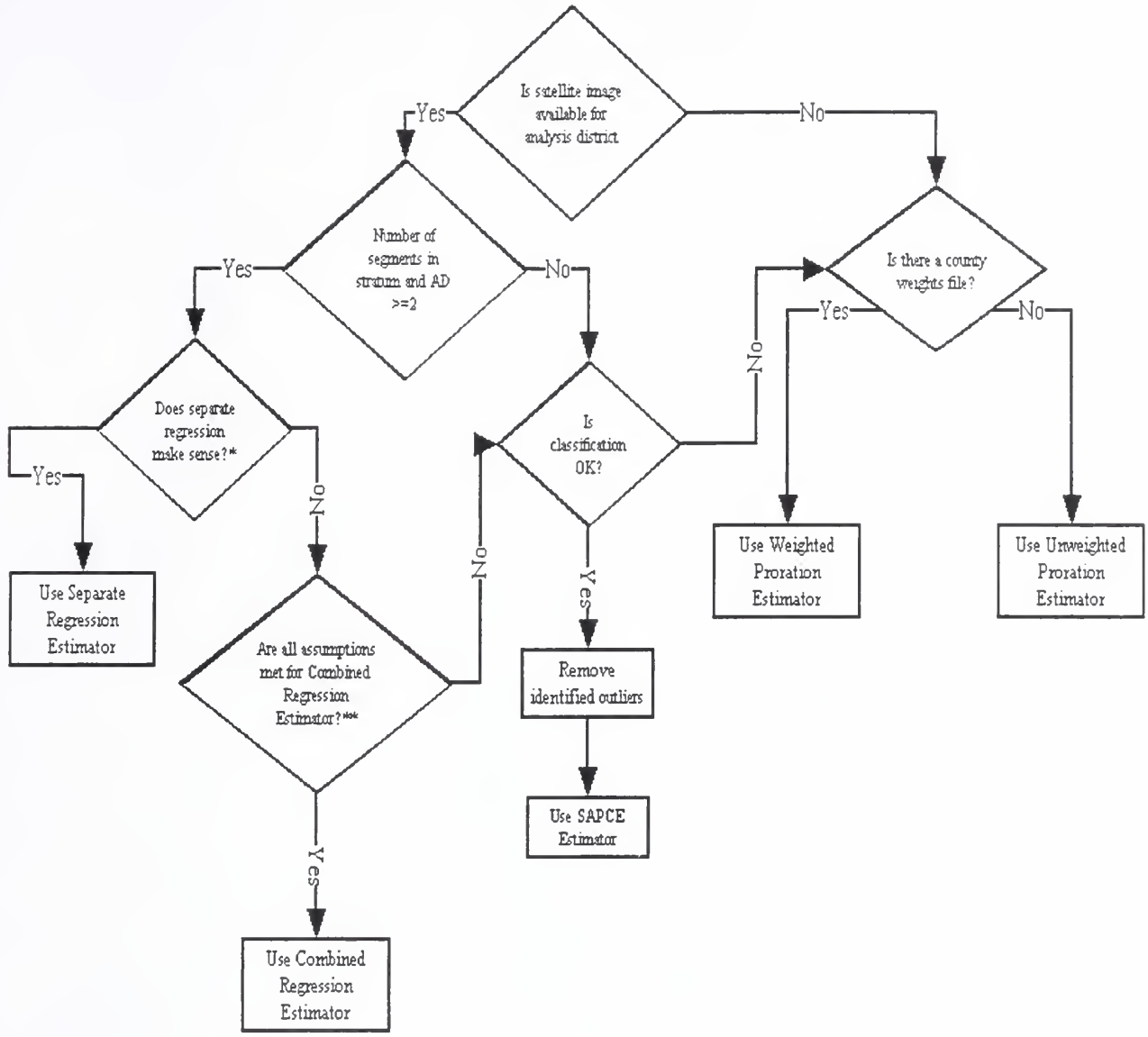
Cochran, W. G., "Sampling Theory When Sampling-Units are of Unequal Sizes," *Journal of the American Statistical Association,* Vol. 37, 1942, pp. 199-212.

Hanuschak, G. A., "Landsat Estimation with Cloud Cover," *Machine Processing of Remotely Sensed Data,* Symposium Proceedings, Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana, 1976.

Walker, G., and Sigman, R., *The Use of LANDSAT for County Estimates of Crop Areas: Evaluation of the Huddleston-Ray and the Battese-Fuller Estimators,* U. S. Department of Agriculture, Statistical Reporting Service, 1982.

Von Steen, D. H., and Wigton, W. H., *Crop Identification and Acreage Measurement Utilizing LANDSAT Imagery,* U. S. Department of Agriculture, Statistical Reporting Service, 1976, particularly pp. 124-127.

APPENDIX 1



*To use the separate regression, you should have 10 segments or more in the ground truth for the stratum. For the regression to "make sense," the value of the coefficient should be close to the size of a pixel (in acres), the R^2 should be reasonably high, outlying observations should be examined and eliminated if judged unreasonable, and a graph of number of pixels classified to a crop vs. acres of that crop reported in the ground truth in each segment should indicate that a linear relationship looks reasonable.

** The combined regression requires a number of assumptions to hold. These assumptions are discussed in the text.